DOCUMENT RESUME

ED 466 491 TM 034 230

AUTHOR	von Davier, Alina A.; Holland, Paul W.; Thayer, Dorothy
TITLE	Population Invariance and Chain versus Post-Stratification Methods for Equating and Test Linking.
INSTITUTION	Educational Testing Service, Princeton, NJ.
PUB DATE	2002-03-07
NOTE	20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1-5, 2002).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Equated Scores; Statistical Analysis
IDENTIFIERS	Chained Equipercentile Equating; Invariance; *Linking Metrics; *Stratification

ABSTRACT

The Non-Equivalent-groups Anchor Test (NEAT) design involves two populations, "P" and "Q," of test takes and makes use of an anchor test to link them. Two observed-score equating methods used for NEAT designs are those based on chain equating and those using the anchor to poststratify the distributions of the two operational test scores to a common population, i.e., Tucker equating and frequency estimation. This paper introduced a method that can be used in the NEAT design to study the population invariance of equating methods. The method is applied to study the relative population invariance of Chain and Post stratification equating methods. It combines self-equating (equating at test to itself) with the root mean square difference (RMSD) measure of the population invariance of test linking methods introduced by N. Dorans and P. Holland (2000). The method is illustrated using data from the Advanced Placement examinations. (Contains 1 table, 2 figures, and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 466 491

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

E. Mingo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

The following paper was prepared for presentation at the March 2002 AERA/NCME program and is the copyrighted material of Educational Testing Service. Limited permission to reproduce reports in their entirety in downloadable PDF format is hereby granted for individual professional reference. Requests for additional permission to reproduce or use all or portions of these papers may be made through the ETS Permissions website by completing the convenient form at: http://www.ets.org/legal/copyright/html .

Population Invariance and Chain versus Post-Stratification Methods for Equating and Test Linking

Alina A. von Davier, Paul W. Holland, and Dorothy Thayer

Educational Testing Service, Princeton, NJ

3/07/02

Abstract: The **N**on-**E**quivalent-groups **A**nchor **T**est (**NEAT**) Design involves two populations, P and Q, of test-takers and make use of an anchor test to link them. Two observed-score equating methods used for NEAT designs are those based on chain equating and those using the anchor to post-stratify the distributions of the two operational test scores to a common population—i.e. Tucker equating and frequency estimation. We introduce a method that can be used in the NEAT design to study the population invariance of equating methods. We then apply this method to study the relative population invariance of Chain and Post stratification equating methods. Our method combines self-equating (equating at test to itself) with the RMSD measure of the population invariance of test linking methods introduced by Dorans and Holland (2000). We illustrate our method using data from the AP Examinations.

TM034230

BEST COPY AVAILABLE

Keywords: test linking, equating, population invariance, Non-Equivalent groups Anchor Test Design, self-equating.

1. Introduction

Test equating methods are used to produce scores that are comparable across different test forms. Weaker forms of test linking often use the same computations as test equating but do not necessarily result in scores that are comparable. One of the primary requirements of equating functions is that they should be population invariant. Because strict population invariance is often impossible to achieve, Dorans and Holland (2000) introduced a measure of the degree to which an equating method is sensitive to the population on which it is computed. The measure compares equating or linking functions computed on different subpopulations with the equating or linking function computed for the whole population. Their discussion is restricted to equating designs that involve only one population (such as the equivalent-groups design and the single group design).

The **N**on-**E**quivalent-groups **A**nchor **T**est (**NEAT**) Design involves *two* populations (usually different test administrations), P and Q, of test-takers and makes use of an anchor test to link them. We also want population invariance to hold for equating functions used in the NEAT design, but there are two populations now, so there can be ambiguity as to which population is the one on which the equating (or linking) is done.

For the NEAT design there are several observed–score equating methods that are used in practice. Two important classes of these methods are those we will call Chain equating and Post-Stratification equating, following Holland (2002).

In this paper we examine the relative population invariance of Chain versus Post-Stratification equating methods in the NEAT Design. We use the existence of two subpopulations, such as Male and Female examinees, to mimic a situation where a test has been reused so that it can be equated to itself and the result compared to the identity function. We use this idea to adapt the Dorans and Holland (2000) measure of Root Mean Square Difference, RMSD(x), to compare the results of Chain and Post-Stratification equating methods. Data from the Advanced Placement program are used to illustrate these ideas.

1.1 The NEAT Design:

Holland (2002) describes the NEAT design. Here we just reiterate some of its basic features. The important idea is the data structure:

	X	V	Y	
P	✓	✓		X, V observed on P
Q		✓	✓	Y, V observed on Q

Usually, X and Y are the "operational tests" given to "test administrations" P and Q, respectively, and V is the "anchor test" given to both P and Q. The anchor test score, V, can be either a part of both X and Y (the internal anchor case) or a separate score (the external anchor case).

The Target Population, T, for the NEAT design is a mixture of P and Q and denoted by $T = wP + (1 - w)Q$. The mixture is determined by a weight w. When $w = 1$, then $T = P$ and when $w = 0$ then $T = Q$. Other choices of w are often used as well.

In this situation we will let the (continuized) cdfs of the score distributions of X, Y and V be denoted by $F(x)$, $G(y)$, and $K(v)$ and will append subscripts as necessary to distinguish between P, Q and the "target population", T.

The two most important test scores, X and Y, are not observed in *both* P and Q, but only one or the other, unlike the anchor test score, V. Thus, assumptions must be made in order to overcome this aspect of the NEAT Design. The different observed score equating methods used in this design each make different assumptions about the distributions of X and Y in the populations where they are not observed.

1.2 Chain and Post-Stratification Equating methods for the NEAT Design

The Chain Equating (CE) and the Post-Stratification Equating (PSE) methods, described in Holland (2002) are two important classes of observed score equating methods used in the NEAT Design. They are briefly described below for the equipercentile case.

A. Chain Equating. Chain equating uses a two-stage transformation of X scores into Y scores. First equate X to V on P and then equate V to Y on Q. These two equating functions are then functionally composed to map X to Y through V. This method is a valid observed score equating method if the following two assumptions, CE1 and CE2, hold, Holland (2002).

CE1: Given any target population T, the link from X to V is population invariant, so that

$$K_P^{-1} \circ F_P(x) = K_T^{-1} \circ F_T(x).$$

(This makes use of the fact that $K_T^{-1} \circ F_T(x)$ is the equipercentile function linking X to V on population T, for any T.)

CE2: Given any target population T, the link from V to Y is population invariant, so that

$$G_Q^{-1} \circ K_Q(v) = G_T^{-1} \circ K_T(v).$$

Applying these two assumptions to the computation of the composed link from X to V to Y, we get

$$e_{XY;T(C)}(x) = G_Q^{-1} \circ K_Q \circ K_P^{-1} \circ F_P(x).$$

See Holland (2002) for more details. We note that because the target population, T, cancels out from the formula for the composed function that equates X to Y, $e_{XY;T(C)}(x)$ is assumed to "work" for any T. In a sense, it is defined to be population invariant, but this is only strictly true for populations that are *mixtures* of P and Q, and not for subpopulations of P or Q.

B. Post-Stratification Equating. This method first estimates the marginal distributions of both X and Y on a target population T (that is a specific mixture of P and Q) and then computes the equipercentile equating function. In order to estimate the distribution of X in Q and the distribution of Y in P, PSE method makes the following assumptions: the conditional distribution of X given V and the conditional distribution of Y given V are population invariant, i.e.,

PSE1: Given a target population T, the conditional distribution of X given V is population invariant, i.e.

$$f_{T(PS)}(x) = \sum_v f_P(x|v) k_T(v),$$

where f(x) denotes the score probabilities for X and k(v) the score probabilities for V.

PSE2: Given a target population T, the conditional distribution of Y given V is population invariant, i.e.

$$g_{T(PS)}(y) = \sum_v g_P(y|v) k_T(v),$$

where g(y) denotes the score probabilities for Y.

Using PSE1 and PSE2, $f_{T(PS)}(x)$ and $g_{T(PS)}(y)$ are computed and from these, continuous cdfs, $F_{T(PS)}(x)$ and $G_{T(PS)}(y)$ are formed. Then Y is equated to X on T through:

$$e_{XY;T(PS)}(x) = G_{T(PS)}^{-1} \circ F_{T(PS)}(x).$$

Note that the equating function, $e_{XY;T(PS)}(x)$, can depend on the choice of T unlike $e_{XY;T(C)}(x)$. Therefore, PSE can be different from CE, though they can also be identical in a particular circumstance that we now discuss.

2. When will CE and PSE both give the same results?

One of the important roles of the anchor test in NEAT Design is to provide information about differences in the *relevant* abilities of the examinees in the two populations, P and Q. This is why the anchor test should be appropriately constructed. Brennan & Kolen (1987) discuss conditions for an appropriate anchor test. Marco, Petersen and Stewart (1983), Petersen, Marco and Stewart (1982), Angoff and Cowell (1985) examined a number of equating methods, with or without an anchor test, varying the similarity of the examinee groups. Other studies focused on matching on the anchor for equating (Lawrence and Dorans, 1990; Livingston, Dorans, & Wright, 1990). These empirical studies are summarized well by the observation:

> "The general [...] finding is that, when the anchor test design is used to equate carefully constructed alternate forms, the groups taking the old and new forms are similar to one another, and the common set is a miniature version of the total test form, then equating methods all tend to give similar results."
>
> (M. Kolen, 1990, pp. 98-99).

The theorem given next is an analytical proof of a part of this statement. More precisely, our first result concerns the case when the anchor test has the *same* distribution on both P and Q (without any additional assumptions about how similar X and Y are or without any further assumption about the anchor test being "a miniature version of the test"). In this situation we show that, as they are described in section 1.2, both CE and PSE will result in exactly the same equating function.

Theorem 1: If, in the NEAT design we have $K_P = K_Q$, then both CE and PSE yield the same equating function and it is

$$e_{XY;T(C)}(x) = e_{XY;T(PS)}(x) = G_Q^{-1} \circ F_P(x).$$

Proof: The case for CE is obvious because the composition, $K_Q \circ K_P^{-1}(x)$ now equals the identity function, so that it cancels out, i.e.,

$$e_{XY;T(C)}(x) = G_Q^{-1} \circ K_Q \circ K_P^{-1} \circ F_P(x) = G_Q^{-1} \circ F_P(x).$$

For the case of PSE, suppose $K_P = K_Q$, then the score probabilities satisfy

$$k_T = wk_P + (1 - w)k_Q = k_P = k_Q, \text{ for any } T = wP + (1 - w)Q.$$

Hence,

$$f_{T(PC)}(x) = \sum_v f_P(x|v)k_T(v) = \sum_v f_P(x|v)k_P(v) = f_P(x), \text{ and}$$
$$g_{T(PC)}(y) = \sum_v g_Q(y|v)k_T(v) = \sum_v g_Q(y|v)k_Q(v) = g_Q(y).$$

Once continuized, we must also have $F_{T(PC)}(x) = F_P(x)$, and $G_{T(PC)}(y) = G_Q(y)$, from which the result for PSE follows. QED

Also note that the theorem will also hold for the Tucker and chain linear equating methods (see Holland, 2002, p. 8, on the relationship between linear and equipercentile equating functions).

This theorem shows that CE and PSE are closely connected when the distributions of V are similar on both P and Q, or, in other words CE and PSE must yield nearly identical results when the two populations are similar in abilities.

The next theorem addresses the case when the distributions of the anchor test can be very different for P and Q, but the anchor test is highly correlated with both X and Y. While the statement of Theorem 2 is fairly obvious we include it because we believe that it lies behind the often stated conclusion that a high correlation between the anchor and the operational test is important.

Theorem 2: If, in the NEAT design we have X = V = Y, so that there is a perfect correlation between the anchor test and the other two tests, then both CE and PSE give the same equating function and it is:

$$e_{XY;T(C)}(x) = e_{XY;T(PS)}(x) = x, \text{ the identity.}$$

Proof: Now, because X = V = Y we have $F_T(x) = G_T(x) = K_T(x)$ for any T. The case for CE follows from the fact that now both $G_Q^{-1} \circ K_Q(x) = x$ and $K_P^{-1} \circ F_P(x) = x$ so that

$$e_{XY;T(C)}(x) = G_Q^{-1} \circ K_Q(K_P^{-1} \circ F_P(x)) = G_Q^{-1} \circ K_Q(x) = x,$$

regardless of how different F_P and F_Q are.

For the case of PSE, note that the conditional score probabilities $f_P(x|v)$, $f_Q(x|v)$, $g_P(y|v)$ and $g_Q(y|v)$ are all 0 unless x = v = y, and then they equal 1. Then the score probabilities satisfy

$$f_{T(PC)}(x) = \sum_v f_P(x|v)k_T(v) = k_T(x) = f_T(x) = wf_P(x) + (1 - w)f_Q(x), \text{ and}$$

$$g_{T(PC)}(y) = \sum_v g_Q(y|v)k_T(v) = k_T(y) = g_T(y) = wg_P(y) + (1 - w)g_Q(y) = f_T(y).$$

Hence the two sets of score probabilities are the same. Once continuized, we must also have $F_{T(PC)}(x) = G_{T(PC)}(x)$ from which the result for PSE follows. QED

Theorem 2 might seem trivial at the first sight, but what it proves analytically that when the anchor is a perfect parallel form of the two tests, then the two methods are population invariant (regardless how big the difference is between the population is). The assumption of perfect parallel tests forms is made, for example, by Levine-observed score equating method.

Theorem 2 will also hold for the Tucker and chain linear equating methods.

We can, therefore, only distinguish between CE and PSE when the distribution of V is *sufficiently different* between P and Q, and the anchor V is *sufficiently different* from X and Y. This is the case we will consider in the rest of this paper.

3. Self-equating in the NEAT Design

"Equating a test to itself" has a long history in test equating (see Petersen, Marco, & Stewart, 1982; Marco, Petersen, & Stewart, 1983; Harris and Crouse, 1993). We shall use it here but in a way that we think is somewhat different from how it has been used in the past. In our use of self-equating, the implicit assumptions that it makes about population invariance combine to yield testable predictions.

In the NEAT Design, a situation where self-equating suggests itself arises when we can subdivide P and Q into two subpopulations, such as Males and Females (M and F). In terms of the data structures of the design this can be seen in the following terms:

	NEAT Design				Reusing X on Q				M and F as P and Q		
	X	V	Y		X	V	X			X	V
P	✓	✓		→	✓	✓		→	M	✓	✓
Q		✓	✓			✓	✓		F	✓	✓

On the left side we have the NEAT design as described earlier. In the middle we have the NEAT design when X is reused on the "new form sample" Q rather than being a new test Y. Finally, on the right side we see that if we treat the two subpopulations, M and F, as two different test administrations (rather than as two subpopulations of the same test administration) then we get the same data structure that arises when X is reused on the "new form sample."

This observation suggests equating X to itself through V (either with CE or PSE) with M and F treated as the two "administrations". This can be done both on P and on Q to see the stability of the results. In the next two sub-sections we examine what happens to CE and PSE when equating X to X in a NEAT Design.

3.1 Self-equating and Chain Equating for the NEAT Design

The CE-assumptions, CE1 and CE2, in self-equating (SE) become:

CE1(SE): $K_P^{-1} \circ F_P(x) = K_T^{-1} \circ F_T(x)$, and therefore $F_{T(C)}(x) = K_T \circ K_P^{-1} \circ F_P(x)$,
CE2(SE): $F_Q^{-1} \circ K_Q(v) = F_T^{-1} \circ K_T(v)$, and therefore $F_{T(C)}^{-1}(x) = F_Q^{-1} \circ K_Q(v) \circ K_T^{-1}$.

Hence: $e_{XX;T(C)}(x) = F_{T(C)}^{-1}(x) \circ F_{T(C)}(x) = x = F_Q^{-1} \circ K_Q(v) \circ K_P^{-1} \circ F_P(x)$,

which is a testable prediction with data for which self-equating is possible. Note that the equation, $x = F_Q^{-1} \circ K_Q(v) \circ K_P^{-1} \circ F_P(x)$, does not depend on T, but does depend on P and Q.

3.2 Self-equating and Post Stratification Equating for the NEAT Design

The PSE-assumptions, PSE1 and PSE2, in self-equating become:

PSE1(SE): $f_{T(PC)}(x) = f_{PT}(x) = \sum_v f_P(x|v)k_T(v)$, which continuizes to $F_{PT}(x)$,

PSE2(SE): $f_{T(PC)}(x) = f_{QT}(x) = \sum_v f_Q(x|v)k_T(v)$, which continuizes to $F_{QT}(x)$.

Hence: $e_{XX;T(PS)}(x) = F_{QT}^{-1} \circ F_{PT}(x) = F_{T(PC)}^{-1} \circ F_{T(PC)}(x) = x$. Note that the expression, $F_{QT}^{-1} \circ F_{PT}(x)$, can depend on T. If we let T = P then we get

$$e_{XX;P(PS)}(x) = F_{QP}^{-1} \circ F_{PP}(x) = F_{QP}^{-1} \circ F_P(x),$$

and if we let T = Q then we get

$$e_{XX;Q(PS)}(x) = F_{QQ}^{-1} \circ F_{PQ}(x) = F_Q^{-1} \circ F_{PQ}(x).$$

Thus we get two different functions, $F_{QP}^{-1} \circ F_P(x)$ and $F_Q^{-1} \circ F_{PQ}(x)$, which should both be the identity function if the population invariance assumptions for PSE hold. Again these are testable predictions.

4. RMSD for Self-Equating using both Chain and Post Stratification

Dorans and Holland (2000) define RMSD(x) as a measure of the degree to which an equating procedure *fails* to be population invariant across a given set of subpopulations of a base population, P. At each X-score, RMSD(x) is the root-mean-square difference between the equating functions computed on each subpopulation and the equating function computed on the whole population, P. It is standardized by dividing by the standard deviation of Y on P, so that it is a type of "effect size".

Holland (2002) describes the RMSD(x) measure when applied to the NEAT Design and gives explicit formulas for CE and PSE.

In our analysis we make use of the natural self-equating that arises when the two subpopulations (say M and F) of the original P of the NEAT Design are treated as if they are themselves two different "test administrations" where X has been reused, as discussed in Section 3.1 and 3.2. In the notation we use here, the two subpopulations, M and F, are called

P and Q, respectively. We will call the original P the target population T. The weights, w_P and w_Q, are, in this interpretation, the weights we give to M and F. In our application we take these to be proportional to the relative sizes of M and F in P, i.e., of P and Q in T.

$$RMSD_{CE}(x) = \sqrt{\frac{w_P(e_{XX;T(C)}(x)-x)^2+w_Q(e_{XX;T(C)}(x)-x)^2}{\sigma^2_{XT}}} = \frac{|e_{XX;T(C)}(x)-x|}{\sigma_{XT}}.$$

$$RMSD_{PSE}(x) = \sqrt{\frac{w_P(e_{XX;P(PS)}(x)-x)^2+w_Q(e_{XX;Q(PS)}(x)-x)^2}{\sigma^2_{XT}}}.$$

The $RMSD_{CE}$ simplifies because the two equating functions, the ones for P and Q, are, as usual for CE, the same. We may compare these two RMSD's because they are both "effect sizes" relative to the same standard deviation. They put the violations of the two "population invariance" assumptions (for CE and PSE) into the same scale. We think this is helpful in deciding which method, CE or PSE, is "more" population invariant, that is, which is *less* dependent on the population used for the equating. Some way of putting the two sets of assumptions, CE1-2 and PSE1-2, on the same footing is necessary because, as stated, these assumptions involve very different distributional quantities for which there is no obvious comparison.

5. An Example Using Data from the AP Examinations

We will use data from the AP Examinations to illustrate our approach to addressing the question: which of the two methods, CE or PSE, is less sensitive to the population invariance assumption. The data discussed here are from the 1998 and 2000 administration of the AP English Language & Composition Examination. In our data sets there were 79,434 examinees in the 1998 administration and 112,868 examinees in the 2000 administration.

We used the Multiple Choice (MC) data from the tests, and did not use the Free Response data at all. This AP test uses a NEAT Design with the year 2000 test being linked back to the one given in 1998. The anchor test is an embedded anchor (EQ) within the MC component of the whole test. The two subpopulations we examined were Males and Females. In 1998 there were 30,217 male and 49, 217 female test takers, and in 2000 there were 42,317 male and 70,551 female test takers.

The effect size computations given in Table 1 show that the M/F differences were about the same size as the test year differences on the

anchor test. We think this observation makes the M/F comparison more relevant to the year to year comparison than if these differences had been less similar.

Table 1: Effect Size calculations for Male/Female differences on the Anchor Test. MC Anchor-Test Data from the 1998 and 2000 Administrations of the AP English Language & Composition Test.

Year	Mean Males	SD Males	Mean Females	SD Females	Mean All	SD All	M-F Means	Effect size
1998	9.408	3.16	9.166	3.15	9.258	3.16	0.242	7.7%
2000	9.152	3.21	8.903	3.23	8.996	3.22	0.249	7.7%

The effect size for the difference between 1998 and 2000 for all examinees is (9.258 – 8.996)/3.19 = 8.2%. (3.19 is the average of 3.22 and 3.16). Thus, the 7.7% effect size for the M/F differences each year are similar to the 8.2% effect size for the difference between the two years.

The correlations between the X = MC98 and anchor test V = EQ and between MC00 and EQ are both about 0.82 in both P = 1998 and Q = 2000. This correlation is considered relatively low for an internal anchor with the test. This low correlation may be relevant to our analysis of the two equating methods. It is known from empirical studies (Livingston et al., 1995), that PSE is more sensitive than CE is to the degree of correlation between the operational and anchor tests.

All of the equipercentile equating functions computed in this analysis made use of the kernel method of equating (Holland and Thayer, 1989; von Davier, Holland, & Thayer, 2002), outlined in appendix A2 of Holland (2002). This procedure included fitting log-linear models to the joint distributions of X and V and of Y and V on both M and F within each of the two years of data. Before computing the equating functions, the cdfs were all made continuous using the automatic bandwidth selection that "post-smoothed" the "teeth" that were observed in these joint distributions of rounded MC formula scores. The resulting continuous cdfs did not have rapidly changing derivatives. As a result of this approach, the resulting RMSD(x) curves are very smooth.

Results: Figures 1 and 2 show the results for the two years of data separately. We used Males and Females to subdivide the P and Q into M and F. Then we used the formulas resulting from self equating for RMSD(x) for both CE and PSE, as derived in Section 4. The curves in each of these of

these figures are the two RMSD(x) curves, one for PSE and one for CE. We put them on the same graph to aid their comparison over the entire score range of X = operational MC test score each year.

(Figures 1 and 2 go about here)

Both figures show very similar results for both CE and PSE. The variation from the identity function is generally greatest in the middle of the score range and reaches an effect size of 9% in 1998 and 5% in 2000. In both years there is some evidence that self-equating does not lead to the identity function for either method, but in each case, in the middle range of scores where most of the examinees are, there is evidence that CE is slightly less sensitive to the choice of population. We will need many more examples using these methods before we could conclude any thing stronger about their relative merits. We believe that these results show that our approach can be used to give a fairly interpretable basis for comparing the two methods on an important criterion for test equating.

6. Summary

In this paper we introduce a new approach for diagnosing the equating methods in the NEAT Design. We describe testable conditions with the data at hand when two subpopulations are identifiable within P and Q. In this way, we can check which method is less sensitive to violations of the population invariance assumption for the given set of data. With experience we believe this approach can give useful information to aid the decision as to which equating method to use in a given situation.

While it is too early to say much about the two methods, CE and PSE, using our approach, with more examples we may be able to quantify their relative sensitivity to the equating population in various situations of interest. We expect that two factors will be important in studies using our approach. First, the correlations of the anchor test with the operational tests. Second, how different the subpopulations are on the anchor test scores. These factors are known to play a role in the NEAT design and our methods provide another way to examine their effects.

We also point out that the combination of the "self-equating" and RMSD(x) might be used for comparisons of other equating and test linking methods with respect to the sub-population differences. This would include both observed and true scores methods. For use with true score methods the formula for RMSD(x) may need some alteration to accommodate test reliability. This is an interesting topic for a future research.

References

Angoff, W. H. and Cowell, W. R. (1985) An Examination of the Assumption that the Equating of Parallel Forms is Population-Independent. Research Report 85-22, Princeton NJ: Educational Testing Service.

Brennan, R. L. and Kolen , M. J. (1987) Some Practical Issues in Equating. *Applied Psychological Measurement*, **11**, 279-290.

von Davier, A. A., Holland, P. W. & Thayer, D. T (2002) *The Kernel Method of Test Equating*. New York: Springer (in preparation).

Dorans, N. J. and Holland, P. W. (2000) Population Invariance and the Equatability of Tests: Basic Theory and the Linear Case. *Journal of Educational Measurement*, **37**, 281-306.

Harris D. J. and Crouse, J. D. (1993) A Study of Criteria Used in Equating. *Applied Measurement in Education*, **6**, 195-240.

Harris D. J. and Kolen M. J. (1986) Effect of Examinee Group on Equating Relationships. *Applied Psychological Measurement*, **10**, 35-43.

Holland, P. W. and Thayer, D. T. (1989) The Kernel Method of Equating Score Distributions. Research Report 89-7, Princeton NJ: Educational Testing Service.

Holland, P. W. and Thayer, D. T. (2000) Univariate and Bivariate Loglinear Models for Discrete Test Score Distributions. *Journal of Educational and Behavioral Statistics*. **25**, 133-183.

Kolen, M. J. (1990) Does Matching in Equating Work? A Discussion. *Applied Measurement in Education*, **3(1)**, 97-104.

Kolen, M. J. and Brennan, R. J. (1995) *Test Equating: Methods and Practices*. New York: Springer.

Lawrence, I. M. and Dorans, N. J. (1990) Effect on Equating Results of Matching Samples on an Anchor Test. *Applied Measurement in Education*, 3(1), 19-36.

Livingston, S.A., Dorans, N. J., and Wright, N. K. (1995) What Combination of Sampling and Equating Methods Works Best? *Applied Measurement in Education*, **3(1)**, 73-95.

Marco, Petersen, and Stewart (1983) A Test of Adequacy of Curvilinear Score Equating Models. In Weiss, D. J. (Ed.), *New Horizons in Testing* (pp. 147-177). New York: Academic.

Petersen, N. S., Marco, G. L., and Stewart, E. E. (1982) A Test of Adequacy of Linear Score Equating Models. In P. W. Holland and D. B. Rubin (Eds.), *Test Equating* (pp. 71-135). New York: Academic.

Figure 1:
RMSD vs MC Score
English Language & Composition 1998

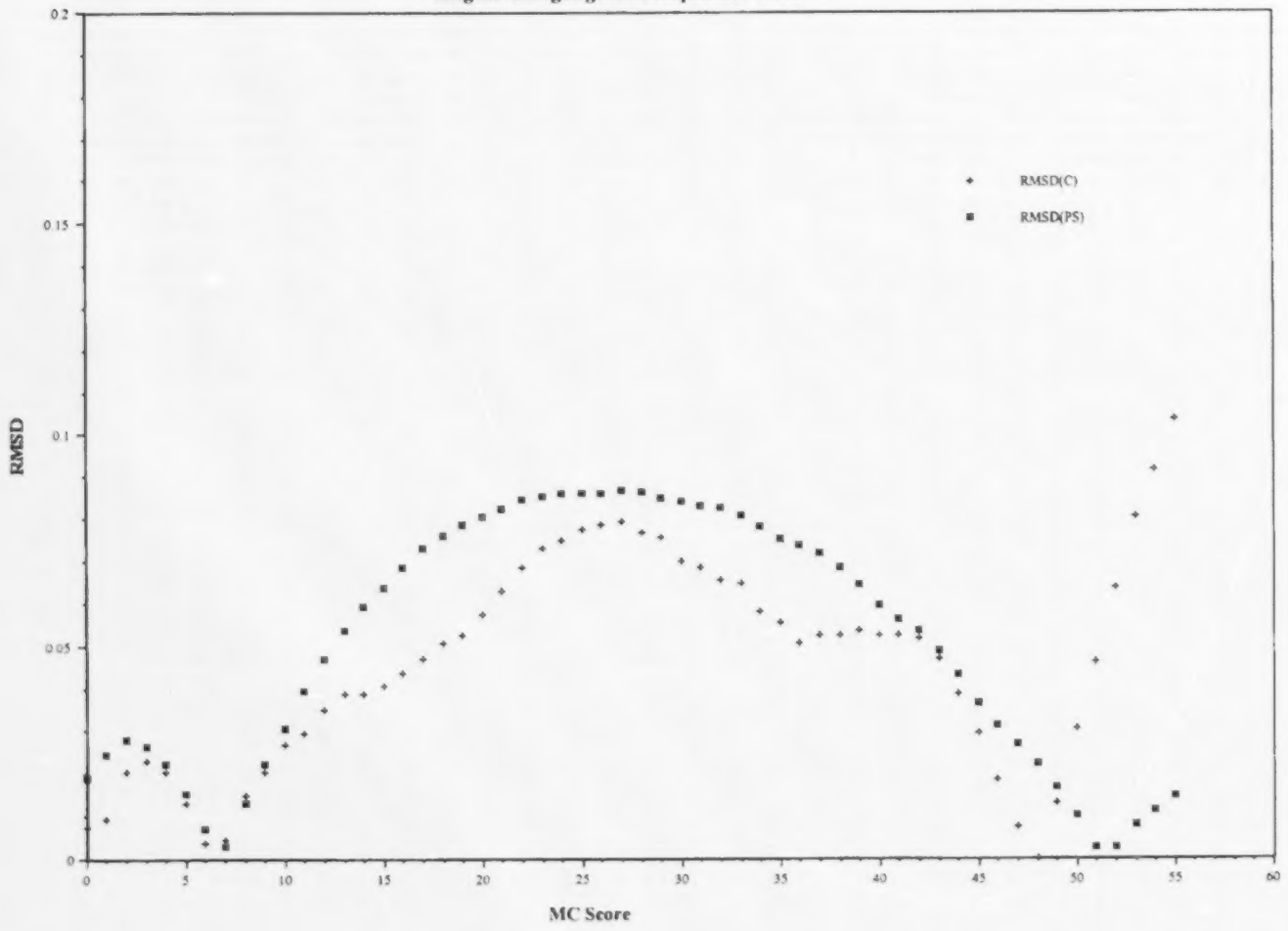
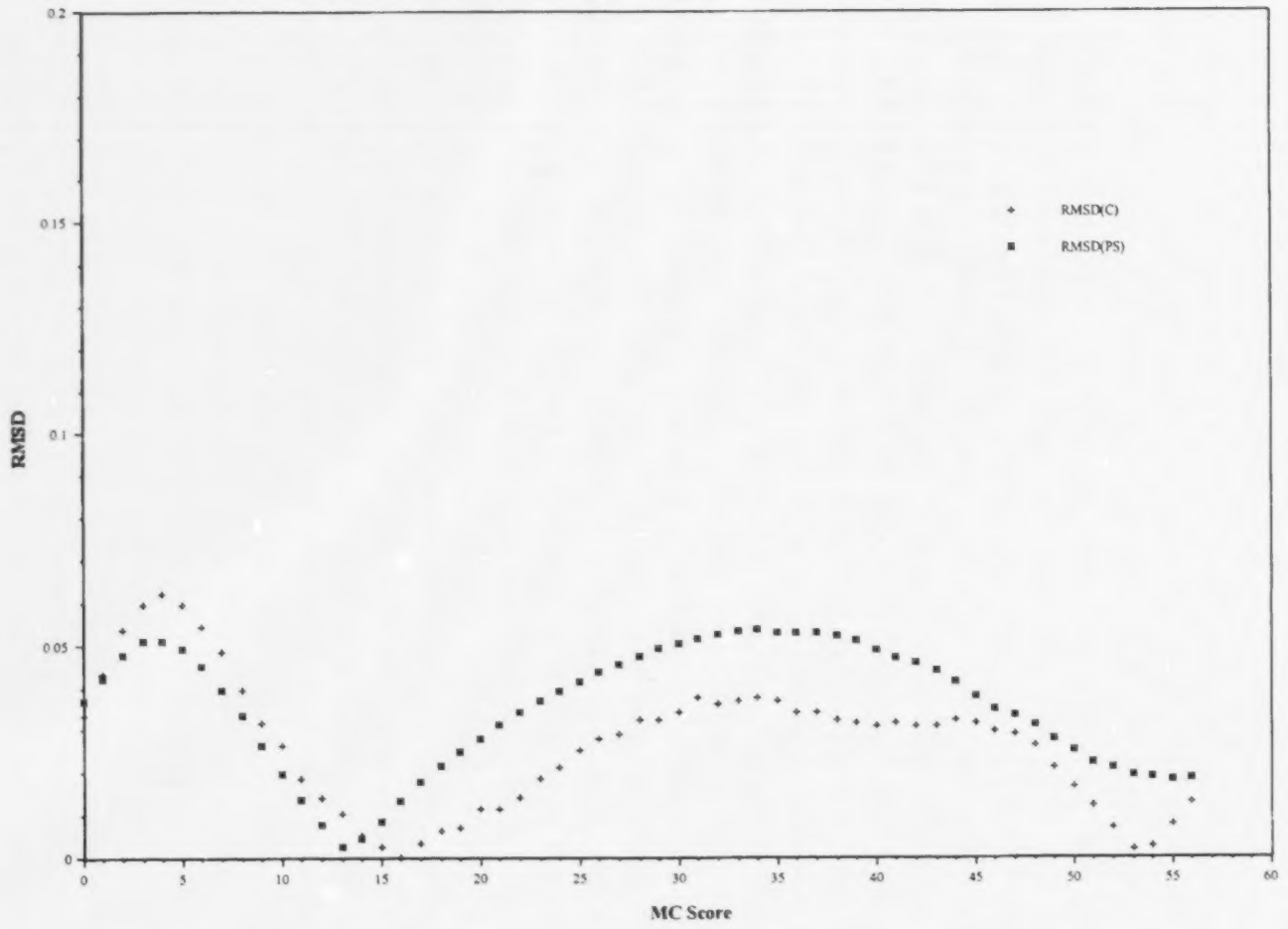


Figure 2:
RMSD vs MC Score
English Language & Composition 2000



19

20

TM034230

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC

NOTICE

Reproduction Basis

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").

EFF-089 (3/2000)